

# Hypothesis Testing Overview



 Uploaded image

what this asking



The question in the image asks about hypothesis testing and the key components of setting up a null hypothesis in statistical contexts. Here's a breakdown of the question:

## 1. Hypothesis Testing:

- The video discusses hypothesis testing as addressing "an idea that can be tested." The end of the video explores the purpose behind creating a null hypothesis.

## 2. Key Factor in Testing Ideas:

- It asks for the key factor that distinguishes ideas that can be statistically tested from those that cannot.

## 3. Criteria for a Good Null Hypothesis:

- The question seeks to identify the "criteria" that define a good null hypothesis.

## 4. Difference Between Null and Alternative Hypotheses:

- It asks for an explanation of the difference between a null hypothesis and an alternative hypothesis in the context of hypothesis testing.

To summarize:

- Define what makes a hypothesis testable.
- Explain the characteristics of a good null hypothesis.
- Clarify the difference between null and alternative hypotheses in hypothesis testing.

Would you like help formulating answers to these questions?



yes



Let's break down each part of the question to provide concise and clear explanations:

## 1. Key Factor that Makes an Idea Testable Statistically:

- The key factor is whether the idea or hypothesis can be measured using empirical data. A hypothesis must be **falsifiable**—meaning there should be a clear, observable, and measurable way to test whether the hypothesis is true or false. If a hypothesis cannot be tested with data or there's no possible way to disprove it, it is not statistically testable.

## 2. Criteria for a Good Null Hypothesis:

- A good null hypothesis (denoted as  $H_0$ ) should:
  1. **Be Clear and Specific:** It must make a clear statement about a population parameter (such as "the mean difference between two groups is zero").
  2. **Testable with Data:** The null hypothesis should be something you can test through observation and statistical analysis.
  3. **Neutral or No Effect:** Typically, the null hypothesis suggests no effect or no relationship (e.g., "there is no difference in means between two groups").
  4. **Use Statistical Parameters:** It often relates to population parameters like means, proportions, or variances.

A null hypothesis must provide a baseline assumption that can be tested and either rejected or not rejected based on the data.

## 3. Difference Between a Null Hypothesis and an Alternative Hypothesis:

- **Null Hypothesis (  $H_0$  ):** This is the hypothesis that suggests there is no effect or no relationship between variables. It is the default assumption in hypothesis testing. For example, "The new drug has no effect on blood pressure compared to a placebo."
- **Alternative Hypothesis (  $H_A$  ):** This is the hypothesis that there **is** an effect or a difference. It represents the statement you want to test for. For example, "The new drug lowers blood pressure compared to a placebo."

In hypothesis testing, you aim to test whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

Would you like more examples or details on any of these explanations?



 Uploaded image

what does this mean



The question in the image is about understanding the distinction between population parameters and sample statistics in hypothesis testing. Here's a breakdown of what it's asking:

### 1. Population vs. Sample:

- The sentence emphasizes that the results of hypothesis tests refer to **population parameters**, not **sample statistics**. In other words, the conclusion we draw from a test is about the overall population, even though we base the test on a sample.

## 2. Terms Explained:

- $x_i$ : Individual sample values or observations. These are the data points collected in a sample.
- $\bar{x}$ : The sample mean, which is the average of the sample values  $x_i$ . It's a statistic that estimates the population mean.
- $\mu$ : The population mean, which is the true average value of the population. This is typically unknown and what we aim to draw conclusions about.
- $\mu_0$ : The hypothesized population mean under the null hypothesis. It represents the value that the population mean is assumed to be in the null hypothesis.

## 3. Explanation of the Sentence:

- The sentence is highlighting that while we calculate statistics like  $\bar{x}$  (sample mean) from the sample data, our conclusions are about  $\mu$  (population mean). Therefore, hypothesis testing uses sample data to infer properties about the entire population.

In essence, the question wants you to clarify the relationship between these terms and how sample statistics ( $x_i$ ,  $\bar{x}$ ) are used to make inferences about population parameters ( $\mu$ ,  $\mu_0$ ) in the context of hypothesis testing. Would you like further clarification on any of these terms?



how to answer this question



To answer this question, you'll want to explain the roles of the sample and population in hypothesis testing, focusing on the terms given. Here's a concise answer you can use or adapt:

In hypothesis testing, we aim to make inferences about the **population** based on data from a **sample**. Here's what each term represents:

- $x_i$ : These are individual sample observations. They are the raw data points we collect in a study.
- $\bar{x}$ : This is the **sample mean**, which is the average of all  $x_i$  values. It serves as an estimate of the population mean.
- $\mu$ : This is the **population mean**, representing the true average for the entire population. It is generally unknown and what we wish to learn about.
- $\mu_0$ : This is the hypothesized population mean under the **null hypothesis**. We use  $\mu_0$  to make a statement about the population mean that we then test with our sample data.

In hypothesis testing, while we calculate  $\bar{x}$  from our sample, our conclusions pertain to  $\mu$ , the population mean. This means that even though the sample mean  $\bar{x}$  is used for calculations, the test ultimately addresses whether  $\mu$  is likely equal to  $\mu_0$  or not.

In summary, sample statistics ( $x_i$  and  $\bar{x}$ ) are used to estimate and draw conclusions about population parameters ( $\mu$  and  $\mu_0$ ). The hypothesis test result is a statement about the population, not just the sample.

This answer ties each term to its role in the hypothesis testing process and reinforces the main idea that hypothesis tests are concerned with population parameters, even though they rely on sample data. Let me know if you need any further adjustments!



that doesnt really answer the questin



To answer the question more directly, let's break it down specifically in relation to the concepts and provide a concise explanation that connects them clearly:

In hypothesis testing, the goal is to make inferences about the **population parameter** (like  $\mu$ ) based on data from a **sample**. This question emphasizes that the outcome of a hypothesis test refers to the population parameter (e.g.,  $\mu$ ) and not just to the sample statistics like  $\bar{x}$ .

Here's what each term means and how they connect to this idea:

1.  $x_i$ : These are the individual data points collected in the sample. They represent specific observations, but they don't describe the overall population.
2.  $\bar{x}$ : This is the **sample mean**, calculated from all  $x_i$  values. It serves as an estimate of the population mean  $\mu$ . However,  $\bar{x}$  only describes the sample itself, not the entire population.
3.  $\mu$ : This is the **population mean**, which represents the true average for the entire population. Hypothesis tests aim to make statements about  $\mu$ , based on the sample data.
4.  $\mu_0$ : This is the hypothesized value of the population mean under the **null hypothesis**. When we conduct a hypothesis test, we are essentially testing whether  $\mu$  is equal to  $\mu_0$ .

The distinction here is that while we use  $\bar{x}$  (from the sample) to estimate and test hypotheses about  $\mu$  (the population mean), the conclusion of the test is about  $\mu$  in relation to  $\mu_0$ , not about  $\bar{x}$ . Thus, we use the sample mean  $\bar{x}$  as evidence to make a decision regarding the population mean  $\mu$  and whether it aligns with  $\mu_0$ .

This explanation directly addresses how each term is related to the overall process of hypothesis testing and how the conclusions we draw are focused on the population parameter  $\mu$ , as opposed to just the sample statistic  $\bar{x}$ .



explain it to a 12 year of old



 **Memory updated**

Alright, let's make it super simple!

Imagine you have a big jar of candies, but you can't see inside. You want to know the *average* number of candies that each jar like this might have. But you can't count every single candy in the world, so you grab a small handful (a **sample**) and count those instead.

Now, here's how each of the terms works in this idea:

1.  $x_i$ : Each candy you count is one **data point**. It's just one piece of the information you collect.
2.  $\bar{x}$ : After you count all the candies in your handful, you find the **average** number of candies per handful. This average tells you about just the candies you picked, not the whole jar.
3.  $\mu$ : This is like the *true* average number of candies in all the jars in the world. You can't actually know it because you haven't checked every jar, but it's what you really want to find out.
4.  $\mu_0$ : This is your guess or starting idea about how many candies you think are in the jar on average. You're testing to see if this guess matches what's really there.

So, even though you only count the candies in your handful (that's  $\bar{x}$ ), you use this information to make a good guess about all the candies in the world (that's  $\mu$ ). The whole test is about seeing if your guess ( $\mu_0$ ) for the world matches up with what you find in your sample.

In the end, the goal is to make a guess about all candies based on the few you counted, even though you don't know exactly what's in every jar!



 Uploaded image



To answer this question, let's break down what it means to "imagine a world where the null hypothesis is true" when calculating a p-value.

## Explanation:

When we calculate a p-value, we start by assuming that the **null hypothesis** is true. This means we're pretending that the situation we're testing is happening in a world where there's no effect, no difference, or whatever the null hypothesis states.

For example, if the null hypothesis is that a coin is fair (50% heads, 50% tails), we first assume that this is true. We then look at our actual data (like flipping the coin 100 times) and ask, "How unusual or extreme is my result, assuming the coin really is fair?"

The **p-value** tells us the probability of getting results as extreme as ours—or more extreme—if the null hypothesis is true. A very low p-value suggests that our observed data would be quite rare in this “null hypothesis world,” making us think, “Hmm, maybe the null hypothesis isn’t true after all.”

In simpler terms, imagining that the null hypothesis is true helps us judge how surprising our actual results are under that assumption. If they’re very surprising, it gives us a reason to doubt the null hypothesis.

This approach focuses on the concept of the p-value as a measure of how much our data agrees with the idea that the null hypothesis is correct. The more surprising the data is under that assumption, the less likely it is that the null hypothesis is true.



what is a p value



A **p-value** is a number that helps us understand how likely our data is, given that the null hypothesis is true. In simple terms, it tells us how “surprising” or “extreme” our results are under the assumption that there is no effect or difference.

Here's how it works:

- **High p-value (like 0.5):** This means our data is quite normal or expected if the null hypothesis is true. There's nothing surprising here, so we have no reason to doubt the null hypothesis.
- **Low p-value (like 0.01):** This means our data is unusual or unexpected if the null hypothesis is true. This gives us a reason to think that the null hypothesis might not be true.

In general, scientists often use a threshold of 0.05. If the p-value is less than 0.05, they usually decide that their results are surprising enough to reject the null hypothesis. If it's above 0.05, they don't reject the null hypothesis.

So, the p-value helps us decide if our data is strange enough to question our original assumption (the null hypothesis).



 Uploaded image



To answer this question, let’s break down why a smaller p-value makes the null hypothesis seem “more ridiculous.”

## Explanation:

A small p-value means that the data we observed is very unlikely if the null hypothesis were true. When we get a tiny p-value, it's as if we're saying, "If the null hypothesis is actually true, then getting this data is so strange or rare that it almost seems impossible."

In other words, a very small p-value indicates that our data doesn't fit well with the assumption that the null hypothesis is true. This makes the null hypothesis look more and more unlikely, or even "ridiculous," because it suggests that we're seeing something rare that doesn't match what we'd expect if the null hypothesis were correct.

So, the smaller the p-value, the more evidence we have that the null hypothesis doesn't make sense given the data we collected. This leads us to doubt the null hypothesis more strongly.

---

This way of thinking connects the p-value directly to how believable (or not) the null hypothesis seems when faced with unusual or surprising data.



 Uploaded image



To answer this question, we'll simulate a p-value based on the information given. Here's a step-by-step explanation:

### 1. Setting up the Hypotheses:

- **Null Hypothesis  $H_0$ :** There's no preference for head tilt direction, so each kiss has a 50% chance of tilting left or right, just like flipping a fair coin.

### 2. Collecting Observed Data:

- Out of 124 couples, 80 tilted their heads to the right. This gives a right tilt rate of 64.5%.

### 3. Calculating the p-value:

- We'll calculate the probability of getting a result as extreme as or more extreme than 64.5% right-tilt by chance, assuming a 50/50 split.

### 4. Using a Binomial Test:

- To simulate this, you'd normally calculate the binomial probability of getting 80 or more right-tilts out of 124, assuming the probability of right-tilt is 0.5.

### 5. Interpreting the p-value:

- Based on the resulting p-value, compare it with the table to determine the evidence level against the null hypothesis.

Let's perform the calculation for you.

The calculated p-value is approximately 0.00078.

## Interpreting the p-value:

According to the table provided:

- This p-value falls between 0.001 and 0.01, which indicates **strong evidence against the null hypothesis**.

In conclusion, based on this p-value, there is strong evidence to suggest that there is a tendency for couples to tilt their heads to the right when kissing, which contradicts the idea of a 50/50 split.



no dont do the calculation give me code so i can do it



Certainly! Here's the Python code you can use to calculate the p-value for this scenario:

python

```
from scipy.stats import binom_test

# Given data
total_couples = 124 # Total number of couples observed
right_tilted = 80    # Number of couples who tilted right
probability_right = 0.5 # Null hypothesis probability of right tilt (50/50 chance)

# Perform the binomial test
p_value = binom_test(right_tilted, total_couples, probability_right,
                     alternative='greater')

# Output the p-value
print("P-value:", p_value)
```

When you run this code, it will calculate the p-value based on the observed data. You can then compare the result with the provided table to determine the strength of evidence against the null hypothesis.



```
ImportError: cannot import name 'binom_test' from 'scipy.stats' (/usr/local/lib/python3.10/dist-packages/scipy/stats/__init__.py)
```



It seems that the function `binom_test` may not be available in the version of SciPy you're using. This function was recently moved to a different module. You can instead use `binomtest` from `scipy.stats`, which is functionally similar. Here's the updated code:



python

```
from scipy.stats import binomtest

# Given data
total_couples = 124 # Total number of couples observed
right_tilted = 80    # Number of couples who tilted right
probability_right = 0.5 # Null hypothesis probability of right tilt (50/50 chance)

# Perform the binomial test
result = binomtest(right_tilted, total_couples, probability_right, alternative='greater')

# Output the p-value
print("P-value:", result.pvalue)
```

This should work with the updated library functions. Let me know if you encounter any other issues!



P-value: 0.0007823670130848343



Great! The p-value you obtained, approximately 0.00078, aligns with the result I calculated earlier.

## Interpreting the Result:

Based on the evidence table:

- Your p-value falls between 0.001 and 0.01, which indicates **strong evidence against the null hypothesis**.

This suggests that it's unlikely the observed head-tilt preference is due to random chance, supporting the idea that there is a genuine preference for tilting right when kissing.



it doesnt fall in that range



You're correct; I misspoke in my interpretation. The p-value you found, **0.00078**, actually falls below **0.001**, indicating **very strong evidence against the null hypothesis** according to the table provided.

This implies that the likelihood of seeing such a strong right-tilt preference by chance, under the assumption of a 50/50 split, is very low. Therefore, you have **very strong evidence** to reject the null hypothesis and suggest that people do have a preference for tilting their heads to the right when kissing.



 Uploaded image

This question addresses whether a p-value can definitively prove or disprove the null hypothesis or establish innocence or guilt.

---

## Explanation:

### 1. Can a Smaller p-value Definitively Prove the Null Hypothesis is False?

- No, a smaller p-value cannot **definitively** prove that the null hypothesis is false. A small p-value indicates that the observed data is unlikely under the null hypothesis, which gives us reason to **reject** it. However, it doesn't provide absolute proof. Hypothesis testing only provides evidence against the null hypothesis, not conclusive proof.

### 2. Can a p-value Prove Innocence or Guilt?

- No, a p-value alone cannot definitively prove innocence or guilt. In statistics, a p-value indicates the strength of evidence against the null hypothesis. In the context of guilt or innocence, it can suggest whether the observed evidence strongly contradicts a null hypothesis (like innocence). However, it does not provide a definitive conclusion—only the likelihood of observing the data if the null hypothesis were true.

### 3. Is There a Specific p-value That Can Definitively Prove One or the Other?

- No p-value, regardless of how low or high, can definitively prove or disprove a hypothesis. A lower p-value (e.g., below 0.05) gives stronger evidence against the null hypothesis, but statistical hypothesis testing inherently deals with probabilities, not certainties. Therefore, p-values are about the strength of evidence, not absolute proof.

In summary, p-values help us decide whether to reject the null hypothesis, but they don't offer definitive proof. They indicate the likelihood of the observed data under the null hypothesis, but they don't make absolute claims about truth, innocence, or guilt.

 Uploaded image

This question is asking you to compare a **one-sided (one-tailed)** test with a **two-sided (two-tailed)** test, and understand how these affect the p-value and interpretation in hypothesis testing.

## Explanation:

### 1. One-Tailed vs. Two-Tailed Tests:

- A **two-tailed test** checks for any difference from the null hypothesis in both directions. For example, if the null hypothesis states that a mean is equal to 10, a two-tailed test checks whether the mean is either significantly higher or lower than 10.
- A **one-tailed test** only looks for a difference in one direction. So, if you're only interested in whether the mean is significantly greater than 10 (and not less than 10), you would use a one-tailed test.

## 2. Changes in Code:

- When switching from a two-tailed test to a one-tailed test, you change the direction of the test. In a two-tailed test, you might use `alternative='two-sided'`. In a one-tailed test, you specify the direction with `alternative='greater'` or `alternative='less'`.
- For example:

```
python

# Two-tailed test
result = binomtest(count, n, p=0.5, alternative='two-sided')

# One-tailed test
result = binomtest(count, n, p=0.5, alternative='greater') # for testing if
greater
```

## 3. Why the p-value is Smaller in a One-Tailed Test:

- A two-tailed test splits the p-value between both tails of the distribution, so it tests for extreme values in both directions. This makes the p-value larger.
- A one-tailed test focuses only on one direction, which concentrates the probability in a single tail, leading to a smaller p-value if the observed result aligns with the test direction.
- Therefore, with the same data, a one-tailed test generally results in a smaller p-value compared to a two-tailed test, which might make it easier to reject the null hypothesis in a one-tailed test.

## Summary:

- In a one-tailed test, you're only checking for evidence in one direction, so the p-value is typically smaller, which might give stronger evidence against the null hypothesis if the data matches the test's direction. However, the choice between one-tailed and two-tailed depends on the research question and whether you're only interested in one specific direction of change.



2. "a test statistic is as or more extreme than the observed test statistic"...

To understand "as or more extreme" we first need to consider formally in terms of the hypothesized population parameter

This is because "as or more extreme" is relative to a hypothesized population parameter which the test statistic estimates

THEN, NEXT, we need to clearly differentiate and compare the "simulated test statistics" from the

```
"observed test statistic"
# "as or more extreme" relative to the hypothesized parameter of the test statistic!
population_parameter_value_under_H0 = 0.5

observed_test_statistic = (patient_data.HealthScoreChange>0).mean()
simulated_test_statistics = IncreaseProportionSimulations_underH0random

SimTestStats_as_or_more_extreme_than_ObsTestStat = \
    abs(simulated_test_statistics - population_parameter_value_under_H0) >= \
    abs(observed_test_statistic - population_parameter_value_under_H0)

print("Which simulated test statistics are \"as or more extreme\"
than the observed test statistic? (of \"\", observed_test_statistic, \"'\", sep="")

pd.DataFrame({'(Simulated) Test Statistic': simulated_test_statistics,
              '>= '+str(observed_test_statistic)+' ?': ['>= '+str(observed_test_statistic)+'
?']*number_of_simulations,
              '"as or more extreme"?': SimTestStats_as_or_more_extreme_than_ObsTestStat})
When the simulation (two code cells back) is based on np.random.seed(1) the output above
includes examples of True for 0.8 >= 0.8 AND
```

0.1 >= 0.8

WTFWTFWTFWTF omgImfao WHY???

WWWWWWHHHHHHHHYYYYYYYYYYYYYYYYYiiiiiiiiiiii!!!!!!!!!!!!!!??????????????

We've got your answer down below.

# figure for demonstration only: code details not of primary concern

```
hist_data =
[IncreaseProportionSimulations_underH0random+np.random.uniform(-0.05,0.05,size=len(IncreasePro
group_labels = ['Bootstrap<br>Sampling<br>Distribution<br>of the<br>Sample<br>Mean<br>
<br>assuming<br>that the<br>H0 null<br>hypothesis<br>IS TRUE']
fig = ff.create_distplot(hist_data, group_labels, curve_type='normal',
                        show_hist=True, show_rug=False, bin_size=0.1)
pv_y = 2.5
pv_y_ = .25
fig.add_shape(type="line", x0=observed_test_statistic, y0=0,
              x1=observed_test_statistic, y1=pv_y,
              line=dict(color="Green", width=4), name="Observed Test Statistic")
fig.add_trace(go.Scatter(x=[observed_test_statistic], y=[pv_y+pv_y_],
                        text=["Observed<br>Test Statistic<br>^"], mode="text", showlegend=False))
# "as or more extreme" also include the "symmetric" observed test statistic...
symmetric_test_statistic = population_parameter_value_under_H0 - \
    abs(observed_test_statistic-population_parameter_value_under_H0)
```

```
fig.add_shape(type="line", x0=symmetric_test_statistic, y0=0,
              x1=symmetric_test_statistic, y1=pv_y,
              line=dict(color="Green", width=4), name="Observed Test Statistic")
fig.add_trace(go.Scatter(x=[symmetric_test_statistic], y=[pv_y+pv_y_],
                        text=["Symmetric" Observed Test Statistic<br>adding for "as or more extreme"
<br>^'], mode="text", showlegend=False))

# Add a transparent rectangle for the lower extreme region
fig.add_shape(type="rect", x0=-0.25, y0=0, x1=symmetric_test_statistic, y1=pv_y,
              fillcolor="LightCoral", opacity=0.5, line_width=0)
# Add a transparent rectangle for the upper extreme region
fig.add_shape(type="rect", x0=observed_test_statistic, y0=0, x1=1.25, y1=pv_y,
              fillcolor="LightCoral", opacity=0.5, line_width=0)

# Update layout
fig.update_layout(
    title="Bootstrapped Sampling Distribution<br>under H0 with p-value regions",
    xaxis_title="Mean Health Score Change", yaxis_title="Density", yaxis=dict(range=[0,
pv_y+2*pv_y_]))
fig.show() # USE `fig.show(renderer="png")` FOR ALL GitHub and MarkUs SUBMISSIONS
Very Clarifying Clarification Questions (if you understand the answer)
```

What is the difference in the "middle value" of the bootstrap sampling distribution of the sample mean VERSUS the bootstrapped sampling distribution of the sample mean under the (assumption of) null hypothesis?

Hint: compare this figure to the initial bootstrapped confidence interval figure of the TUT notebook

What's the difference between the bootstrapped confidence intervals "interval" compared to the "as or more extreme" regions which p-values are based on?

So does the p-value number below match the figure above (of course allowing for the nuance that the figure above is on a "density" rather than counts scale)?

is this what you need



It looks like you're asking for help understanding and explaining code and statistical concepts related to the idea of "as or more extreme" test statistics within the context of hypothesis testing and bootstrap sampling. Let's break down the main concepts and questions posed in your text:

## Key Concepts Explained

### 1. "As or More Extreme" Relative to the Hypothesized Population Parameter:

- In hypothesis testing, "as or more extreme" refers to how far the observed data is from the hypothesized value under the null hypothesis. Specifically, it looks at how likely it is to see data as far or farther from the null hypothesis value in either direction.
- This is typically done by comparing the observed test statistic to a distribution of simulated test statistics under the null hypothesis.

## 2. Observed vs. Simulated Test Statistics:

- **Observed Test Statistic:** This is the value calculated from your actual sample data (e.g., mean health score change in your example).
- **Simulated Test Statistics:** These are values generated under the assumption that the null hypothesis is true, often by repeating the sampling process many times to create a distribution.
- Comparing the observed test statistic to the distribution of simulated test statistics allows you to determine how extreme your observed result is within the context of the null hypothesis.

## 3. Code Explanation:

- In your code, you are computing which simulated test statistics are as or more



no im asking if for question 7 if this si the code its referring to



Yes, this code appears to align with what Question 7 is referring to. Here's why it matches the requirements of Question 7:

### 1. One-Tailed vs. Two-Tailed:

- The question asks about "one-sided" (or "one-tailed") tests versus "two-sided" (or "two-tailed") tests.
- This code focuses on determining which simulated test statistics are as or more extreme than the observed test statistic in both directions (as evidenced by the use of absolute values and the symmetric test statistic). This indicates a two-tailed approach since it considers extremes on both sides of the hypothesized parameter.

### 2. Simulating Extremes:

- The line `SimTestStats_as_or_more_extreme_than_ObsTestStat`` compares each simulated statistic to the observed one, relative to the hypothesized population parameter (0.5). This is consistent with calculating p-values for both sides (two-tailed).
- The code checks whether the simulated test statistics fall within the "as or more extreme" regions on either side, which would be a consideration for two-tailed tests.

### 3. Visual Representation:

- The code block for generating the figure shows two extreme regions (left and right) around the hypothesized mean (0.5), again indicating a two-tailed approach.

## In Summary:

This code calculates and visually represents the p-value by determining how extreme the observed test statistic is in relation to simulated values, taking into account both tails. Therefore, it reflects the two-sided (or two-tailed) test approach discussed in Question 7. If you want to adjust this for a one-

tailed test, you would focus on just one direction (greater or less than) without the symmetry consideration.



answer the question with this new info



Certainly! Based on the code you provided, here's how it connects to Question 7's requirements about one-tailed versus two-tailed tests:

## Explanation for Question 7:

The code you shared focuses on calculating the p-value by comparing the **observed test statistic** to a distribution of **simulated test statistics** under the null hypothesis. This code specifically evaluates how many of the simulated statistics are "as or more extreme" than the observed one in both directions, which reflects a **two-tailed test**.

### 1. Differences in Interpretation (One-Tailed vs. Two-Tailed):

- A **two-tailed test** considers extremes on both sides of the hypothesized parameter. In the code, this is done by looking at both values greater than and less than the observed test statistic relative to the hypothesized value of 0.5. This checks for differences in either direction.
- If we were conducting a **one-tailed test**, we would only check one side. For example, if you were only interested in testing whether the health score change is significantly **greater than** zero, you would only evaluate the right side (values greater than the hypothesized parameter). This means you'd compare values to see if they exceed the observed statistic without considering the symmetric lower extreme.

### 2. Changes in Code for a One-Tailed Test:

- To adapt this code for a one-tailed test, you would adjust the line that defines ``SimTestStats_as_or_more_extreme_than_ObsTestStat`` to only consider one direction. For example:

```
python
```

```
# For a one-tailed test where we're only interested in "greater than":  
SimTestStats_as_or_more_extreme_than_ObsTestStat = \  
    simulated_test_statistics >= observed_test_statistic
```

- This code would now only count simulated values that are greater than or equal to the observed statistic, focusing solely on one side of the distribution.

### 3. Expectations for the p-value:

- In a **two-tailed test**, the p-value will generally be larger because it accounts for extremes on both sides. This code, as it is, captures that broader range of extreme values.

- In a **one-tailed test**, since we only consider one direction, the p-value would typically be smaller, making it easier to find significance in that direction if there is a difference.

## Conclusion:

The provided code is indeed tailored for a **two-tailed test**, as it assesses extremes relative to both tails of the hypothesized parameter. If you were to adjust it for a **one-tailed test**, the p-value would be expected to be smaller, as only one direction (greater than or less than) would be considered. This makes one-tailed tests potentially more sensitive when you have a specific directional hypothesis in mind.



### Demo II (of Hypothesis Testing using p-values) [30 minutes]

#### Further Guidance

The above illustrates rejecting a null hypothesis on the basis of a bootstrapped confidence interval at a 95% confidence level (since the interval "does not cover 0")

This is an ideal way to address hypothesis testing, but it's (unfortunately) also quite common to give "evidence against" a null hypothesis in the form of a p-value

A p-value is the probability that a test statistic is as or more extreme than the observed test statistic if the null hypothesis is true

To understand what the definition of a p-value means, let's consider the definition in reverse

What is the meaning of "if the null hypothesis was true"?

What is the meaning of "a test statistic is as or more extreme than the observed test statistic"?

What is the meaning of "the probability that a test statistic is..."?

1. "if the null hypothesis is true"...

implies that improvements or reductions between FinalHealthScore and InitialHealthScore in individual observations are actually really just "random"

We could therefore just simulate sampling distribution of the "proportion of cases that improved" under the assumption of the null hypothesis that the signs of the differences between InitialHealthScore and FinalHealthScore is actually really just as random as the process of flipping a fair coin.

We'll therefore use the following slightly different version null hypothesis

#### Further Guidance

We're changing

for two reasons; first, this is the version that we want to introduce and construct the foundation of hypothesis testing with; and, second, for a "technical" reason this null hypothesis is also more amenable to the simulation approaches that we're leveraging in STA130.

After seeing how we can use simulation to address using formal hypothesis testing, a very good challenge for students for confirming understanding



would be to determine how bootstrapping could be used to estimate the "chance the vaccine improves patient health" through a confidence interval.

# Do you get the idea here?

# Can you see what's changing in the output below??

```
print(pd.DataFrame({'HealthScoreChange': patient_data['HealthScoreChange'],
                    '> 0 ?': patient_data['HealthScoreChange']>0}))
```

```
random_difference_sign = np.random.choice([-1, 1], size=len(patient_data))
```

```
pd.DataFrame({'HealthScoreChange':
random_difference_sign*patient_data['HealthScoreChange'].abs(),
              '> 0 ?': (random_difference_sign*patient_data['HealthScoreChange'])>0})
```

# And then can you see what's happening here???

```
np.random.seed(1) # make simulation reproducible
```

```
number_of_simulations = 10000 # experiment with this... what does this do?
```

```
n_size = len(patient_data) # 10
```

```
IncreaseProportionSimulations_underH0random = np.zeros(number_of_simulations)
```

```
# generate "random improvement" proportions assuming H0 (vaccine has no average effect) is
true
```

```
# meaning that the "before and after" differences are positive or negative at "random"
```

```
for i in range(number_of_simulations):
```

```
    # why is this equivalent to the suggested idea above?
```

```
    random_improvement = np.random.choice([0,1], size=len(patient_data), replace=True) # <<<
`replace=True` ^^^
```

```
    # why is .mean() a proportion?
```

```
    IncreaseProportionSimulations_underH0random[i] = random_improvement.mean()
```

```
    # why is this the statistic we're interested in? Hint: next section...
```

2. "a test statistic is as or more extreme than the observed test statistic"...

To understand "as or more extreme" we first need to consider

formally in terms of the hypothesized population parameter

This is because "as or more extreme" is relative to a hypothesized population parameter which the test statistic estimates

THEN, NEXT, we need to clearly differentiate and compare the "simulated test statistics" from the "observed test statistic"

```
# "as or more extreme" relative to the hypothesized parameter of the test statistic!
```

```
population_parameter_value_under_H0 = 0.5
```

```
observed_test_statistic = (patient_data.HealthScoreChange>0).mean()
```

```
simulated_test_statistics = IncreaseProportionSimulations_underH0random
```

```
SimTestStats_as_or_more_extreme_than_ObsTestStat = \
```

```
abs(simulated_test_statistics - population_parameter_value_under_H0) >= \
```

```
abs(observed_test_statistic - population_parameter_value_under_H0)
```

```
print("Which simulated test statistics are "as or more extreme"
than the observed test statistic? (of '", observed_test_statistic, ')', sep="")
```

```
pd.DataFrame({'(Simulated) Test Statistic': simulated_test_statistics,
             '>= '+str(observed_test_statistic)+' ?': ['>= '+str(observed_test_statistic)+'
?']*number_of_simulations,
             '"as or more extreme"?': SimTestStats_as_or_more_extreme_than_ObsTestStat})
```

When the simulation (two code cells back) is based on `np.random.seed(1)` the output above includes examples of True for `0.8 >= 0.8` AND

`0.1 >= 0.8`

WTFWTFWTFWTF omgImfao WHY???

WWWWWWHHHHHHHHYYYYYYYYYYYYYYYYYiiiiiiiiiiiiiiii!!!!!!!!!!!!!!??????????????

We've got your answer down below.

# figure for demonstration only: code details not of primary concern

```
hist_data =
[IncreaseProportionSimulations_underH0random+np.random.uniform(-0.05,0.05,size=len(IncreasePro
group_labels = ['Bootstrap<br>Sampling<br>Distribution<br>of the<br>Sample<br>Mean<br>
<br>assuming<br>that the<br>H0 null<br>hypothesis<br>IS TRUE']
fig = ff.create_distplot(hist_data, group_labels, curve_type='normal',
                        show_hist=True, show_rug=False, bin_size=0.1)
pv_y = 2.5
pv_y_ = .25
fig.add_shape(type="line", x0=observed_test_statistic, y0=0,
              x1=observed_test_statistic, y1=pv_y,
              line=dict(color="Green", width=4), name="Observed Test Statistic")
fig.add_trace(go.Scatter(x=[observed_test_statistic], y=[pv_y+pv_y_],
                        text=["Observed<br>Test Statistic<br>^"], mode="text", showlegend=False))
# "as or more extreme" also include the "symmetric" observed test statistic...
symmetric_test_statistic = population_parameter_value_under_H0 - \
                        abs(observed_test_statistic-population_parameter_value_under_H0)
fig.add_shape(type="line", x0=symmetric_test_statistic, y0=0,
              x1=symmetric_test_statistic, y1=pv_y,
              line=dict(color="Green", width=4), name="Observed Test Statistic")
fig.add_trace(go.Scatter(x=[symmetric_test_statistic], y=[pv_y+pv_y_],
                        text=["Symmetric" Observed Test Statistic<br>addrdssing for "as or more extreme"
<br>^"], mode="text", showlegend=False))

# Add a transparent rectangle for the lower extreme region
fig.add_shape(type="rect", x0=-0.25, y0=0, x1=symmetric_test_statistic, y1=pv_y,
              fillcolor="LightCoral", opacity=0.5, line_width=0)
```

```
# Add a transparent rectangle for the upper extreme region
fig.add_shape(type="rect", x0=observed_test_statistic, y0=0, x1=1.25, y1=pv_y,
              fillcolor="LightCoral", opacity=0.5, line_width=0)

# Update layout
fig.update_layout(
    title="Bootstrapped Sampling Distribution<br>under H0 with p-value regions",
    xaxis_title="Mean Health Score Change", yaxis_title="Density", yaxis=dict(range=[0,
pv_y+2*pv_y]))
fig.show() # USE `fig.show(renderer="png")` FOR ALL GitHub and MarkUs SUBMISSIONS
Very Clarifying Clarification Questions (if you understand the answer)
```

What is the difference in the "middle value" of the bootstrap sampling distribution of the sample mean VERSUS the bootstrapped sampling distribution of the sample mean under the (assumption of) null hypothesis?

Hint: compare this figure to the initial bootstrapped confidence interval figure of the TUT notebook

What's the difference between the bootstrapped confidence intervals "interval" compared to the "as or more extreme" regions which p-values are based on?

So does the p-value number below match the figure above (of course allowing for the nuance that the figure above is on a "density" rather than counts scale)?

heres the whole thing



It looks like you've provided detailed guidance and code for using bootstrapping and simulation to calculate p-values and understand hypothesis testing. Let's break down how to approach this within the context of the question:

## Step-by-Step Explanation:

### 1. Understanding "If the Null Hypothesis is True":

- In this context, the null hypothesis states that there is no actual effect on health scores due to the treatment (e.g., a vaccine). Therefore, any observed improvements or reductions in health scores are assumed to be due to random chance.
- The simulation creates a scenario where any change in health scores is as random as a coin flip, indicating no real effect, which aligns with the null hypothesis.

### 2. Simulating Under the Null Hypothesis:

- The code uses random signs for the change in health scores to simulate a scenario where there is no consistent improvement or decline, which matches the null hypothesis that there is no real effect.
- Bootstrapping is then used to simulate the distribution



answer question 7



To answer Question 7 regarding the difference between one-tailed and two-tailed tests and the impact on the p-value:

---

## Explanation for Question 7:

The code you shared focuses on a **two-tailed test** approach by checking "as or more extreme" values on both sides of the hypothesized population parameter, specifically regarding a health score change under the null hypothesis.

### 1. One-Tailed vs. Two-Tailed Tests:

- In a **two-tailed test**, we are interested in any significant difference from the hypothesized parameter in both directions. This means we check for results that are either significantly greater or significantly less than the hypothesized value.
- In a **one-tailed test**, we only check for an effect in one direction (either greater than or less than). This is relevant if we only care about improvement or only about decline, not both.

### 2. How the Code Would Change for One-Tailed vs. Two-Tailed:

- In your code, `SimTestStats_as_or_more_extreme_than_ObsTestStat` checks for extremeness on both sides, which is two-tailed.
- For a **one-tailed test**, you would modify the code to only consider values greater than (or less than) the observed test statistic without the absolute value. For example:

```
python
```

```
# For a one-tailed test where we're only interested in "greater than":
SimTestStats_as_or_more_extreme_than_ObsTestStat = \
    simulated_test_statistics >= observed_test_statistic
```

- This would reduce the p-value because it only considers one side, making it easier to reject the null hypothesis if there is evidence in the specified direction.

### 3. Effect on the p-value:

- In a two-tailed test, the p-value is generally larger because it includes the probabilities of getting results as extreme in both directions.
- In a one-tailed test, the p-value would be smaller for the same observed test statistic because it's focused only on one side, making it more sensitive to detect significance in that direction.

## Conclusion:

The code you've provided aligns with a **two-tailed test** as it evaluates both directions of "extremeness." For a **one-tailed test**, you would adjust the code to focus on just one direction, which would lower the p-value and potentially provide stronger evidence against the null hypothesis if the data supports that single direction.



## 8. Complete the following assignment.

### Fisher's Tea Experiment

#### Overview

A most beloved piece of statistical lore about the (most famous) statistician Ronald Fisher involves cups of tea with milk. Fisher and his friend and colleague, Dr. Muriel Bristol, worked at Cambridge in the 1920s and regularly had tea together. During one of their afternoon tea times, Bristol refused a cup of tea from Fisher because he put milk in first BEFORE pouring in the tea. Bristol said she could taste the difference, and much preferred the taste of tea when the milk was poured in afterward the tea. Fisher didn't think that there could be a difference and proposed a hypothesis test to examine the situation.

Fisher made 8 cups of tea, 4 with milk added in first and 4 with tea added in first, and gave them to Dr. Bristol without her seeing how they were made and she would say if she thought the tea or the milk was poured first. As it turned out, Bristol correctly identified if the tea or milk was poured first for all 8 of the cups. Fisher, being a skeptical statistician wanted to test if this could be happening by chance with Bristol just randomly guessing (or whether there was evidence against an assumption of Bristol just randomly guessing), and subsequently designed a statistical hypothesis test to do so.

Suppose you run an experiment like this with students in STA130. You get a random sample of 80 STA130 students to each taste one cup of tea and tell you whether they think the milk or tea was poured first. Suppose 49 students are able to correctly state which was poured first. Provide a statistical analysis of this experiment as guided through the following set of questions.

#### Data

49 out of a sample of 80 students are able to correctly state which was poured first.

#### Deliverables

While you can choose how to approach the project, we are interested in evaluating your report relative to the following deliverables:

Clarity of your documentation, code, and written report

Description of the population (and sample) and parameter of interest (and corresponding observed test statistic)

Formal null hypotheses

Provide a formal version

based on the population parameter

Provide an informal interpretive statement explaining

in more casual everyday common language

Alternative hypothesis

in terms of

Quantitative analysis addressing the validity of

Explanation of the method clearly articulating the purpose of the usage of statistic(s) to address the population parameter of interest

Comments

Regarding the population (and the sample), there is a clear difference between the experiment

with STA130 students considered here and the original motivating experimental context of Fisher and Bristol.

the sample size is different.

but so too is the nature of the population. the parameter in question might be considered more personalized in the original experiment; whereas, the parameter in the context of STA130 students might be a more abstract concept

The analysis here could be approached from the perspective of formal hypothesis testing, which would likely involve the simulation of a sampling distribution under in order to estimate p-value with respect to the null hypothesis based on the observed test statistic (how?), concluding with the assessment of based on an interpretation of the meaning of the p-value relative to but a confidence interval approach to considering the hypothesis could also be considered. Consider organizing your report within the following outline template.

Problem Introduction

Relationship between this experiment and the original with Fisher and Bristol

Statements of the Null Hypothesis and Alternative hypothesis

Quantitative Analysis

Methodology Code and Explanations

(if needed/optional) Supporting Visualizations

Findings and Discussion

Conclusion regarding the Null Hypothesis

Further Instructions:

When using random functions, you should make your analysis reproducible by using the

`np.random.seed()` function

whats this asking



This assignment is asking you to conduct a **statistical analysis** of an experiment similar to Fisher's famous tea-tasting experiment. The goal is to evaluate whether the observed data (49 out of 80 students correctly identifying the pouring order) could have occurred by **random guessing** or if there's evidence against that, similar to Fisher's experiment with Dr. Bristol.

Here's a breakdown of what the assignment is asking for:

## 1. Clarity of Report and Code:

- You're expected to clearly document your code and explain your approach.
- Provide a clean, well-organized analysis with explanations for every step you take in the statistical testing process.

## 2. Description of Population and Parameter of Interest:

- **Population:** STA130 students (those participating in the experiment).
- **Sample:** The 80 students you took for the experiment.
- **Parameter of Interest:** The proportion of students who can correctly identify whether milk or tea was poured first. This corresponds to the probability of correctly guessing, which in the null hypothesis would be 50% (random guessing).

- **Observed Test Statistic:** The actual proportion of correct guesses (49 out of 80 students = 0.6125).

### 3. Formal Null and Alternative Hypotheses:

- **Null Hypothesis  $H_0$ :** The students are guessing randomly, meaning the proportion of correct guesses is 0.5 (50% chance).
  - **Informal Statement:** "Students can't really tell the difference, and any correct answers are just due to chance."
- **Alternative Hypothesis  $H_A$ :** The students can identify whether the milk or tea was poured first with a probability higher than 0.5.
  - **Informal Statement:** "Students are able to correctly identify the pouring order better than by random chance."

### 4. Quantitative Analysis and Methodology:

- **Purpose:** You are performing a **hypothesis test** to determine whether the students' success rate (49/80) is significantly better than random guessing.
- **Approach:** Likely using a **binomial test** or a **simulation-based approach** to calculate a p-value and determine whether the observed proportion of correct guesses is statistically significant.
  - You might also construct a **confidence interval** around the observed success rate to see if it includes 0.5, which would indicate that the result could be due to chance.

### 5. Simulation of a Sampling Distribution Under the Null Hypothesis:

- You could use a simulation to estimate how often a sample of 80 students would get 49 or more correct answers if they were guessing randomly (null hypothesis:  $p = 0.5$ ).
- This would involve generating many random samples (simulations) where each student has a 50% chance of guessing correctly, and seeing how extreme the actual result (49 correct) is compared to these random samples.

### 6. Conclusion Regarding the Null Hypothesis:

- Based on your analysis (whether through a p-value or confidence interval), conclude whether or not you reject the null hypothesis.
- If the p-value is small (typically less than 0.05), you would reject the null hypothesis and conclude that students are not just guessing.
- If the p-value is large, you fail to reject the null hypothesis, meaning the data does not provide strong evidence that students can tell the difference beyond chance.

### Report Outline Template:

1. **Problem Introduction:** Briefly introduce the experiment and the question being tested.
2. **Relationship to Original Experiment:** Compare this student experiment to Fisher's original tea experiment with Dr. Bristol.
3. **Null and Alternative Hypotheses:** Clearly state the null and alternative hypotheses, both formally and informally.
4. **Quantitative Analysis:** Perform the hypothesis test, using either a simulation, a binomial test, or confidence intervals.
5. **Methodology and Code:** Show and explain the code used for your analysis (e.g., binomial test, simulation, etc.).



6. **Supporting Visualizations (Optional):** You can include graphs like histograms of the simulated results, confidence interval plots, etc.
7. **Findings and Discussion:** Discuss your results and what they mean in terms of the hypotheses.
8. **Conclusion:** Conclude by stating whether you reject or fail to reject the null hypothesis.

## Next Steps:

- You need to write the code for the hypothesis test (binomial or simulation) to compute the p-value, based on the null hypothesis that students are guessing.
- Provide clear documentation and explanations for each step in the analysis. If you'd like help with coding any part of this, feel free to ask!



ok well do the assignment



Let's go step-by-step and complete the assignment.

## 1. Problem Introduction

We are conducting an experiment similar to Ronald Fisher's famous tea-tasting experiment. In this case, 80 STA130 students were asked to determine whether the milk or tea was poured first in a cup. Out of 80 students, 49 correctly identified the order. We want to determine whether this result could have happened by random guessing, or whether there is evidence that the students are better than chance at identifying the pouring order.

## 2. Relationship to Fisher and Bristol's Experiment

Fisher's original experiment with Dr. Muriel Bristol involved identifying whether tea or milk was poured first in a cup, and Dr. Bristol correctly identified the order for all 8 cups. Fisher's hypothesis test was designed to determine whether this result could have happened by random guessing or whether Dr. Bristol could genuinely tell the difference.

In this STA130 experiment, the sample size is larger (80 students), but the nature of the hypothesis test is similar. We are testing whether the success rate of 49 out of 80 students is statistically significantly better than random guessing.

## 3. Null and Alternative Hypotheses

- **Null Hypothesis  $H_0$ :** The students are guessing randomly, meaning that the probability of correctly identifying the pouring order is 0.5 (50%).

$$H_0 : p = 0.5$$



- **Informal Statement:** The students cannot tell the difference between the two methods, and any correct answers are due to chance.
- **Alternative Hypothesis  $H_A$ :** The students can identify the correct pouring order better than random guessing, meaning the probability of a correct guess is greater than 0.5.

$$H_A : p > 0.5$$

- **Informal Statement:** The students can correctly identify the pouring order better than by random guessing.

## 4. Quantitative Analysis

We will use a **binomial test** to assess whether the result of 49 out of 80 correct identifications is statistically significant. Under the null hypothesis, we expect a 50% success rate due to random guessing. We will calculate the p-value to determine the likelihood of observing 49 or more correct answers by chance.

Let's perform the analysis.

## 5. Methodology and Code

### Step 1: Set up the data

- Total number of students: 80
- Correct guesses: 49
- Probability of guessing correctly under the null hypothesis: 0.5

### Step 2: Perform a binomial test

python

```
from scipy.stats import binomtest

# Given data
total_students = 80
correct_guesses = 49
probability_under_null = 0.5

# Perform the binomial test
result = binomtest(correct_guesses, total_students, probability_under_null,
                    alternative='greater')

# Output the p-value and test statistic
print("P-value:", result.pvalue)
```

Let me know if you would like me to execute this for you, or you can run the code in your own environment.

## 6. Supporting Visualizations (Optional)

You could create a visual representation of the binomial distribution and highlight where the observed test statistic (49/80) falls relative to the expected distribution under the null hypothesis.

For example, a histogram of simulated results could show how often we'd expect 49 or more correct guesses under the null hypothesis of random guessing.

## 7. Findings and Discussion

Based on the results of the binomial test:

- If the **p-value** is small (typically less than 0.05), this means that the observed result (49 correct guesses) is unlikely to occur by random chance alone. In this case, we would reject the null hypothesis and conclude that the students can identify the pouring order better than by random guessing.
- If the **p-value** is large, we fail to reject the null hypothesis, meaning there is not enough evidence to suggest that students can do better than random guessing.

## 8. Conclusion

Depending on the p-value from the binomial test, we will either reject or fail to reject the null hypothesis:

- **Reject the null hypothesis:** If the p-value is small, we conclude that there is evidence suggesting the students are better than random guessing.
- **Fail to reject the null hypothesis:** If the p-value is large, we conclude that the data does not provide enough evidence to say the students can identify the pouring order better than chance.

Would you like me to run the code to get the p-value and finalize the conclusion?



**\*\*Alternative Hypothesis  $H_A \setminus A_H$ \*\*:** The students can identify the correct pouring order better than random guessing, meaning the probability of a correct guess is greater than 0.5.  
 $H_A: p > 0.5$   
 $H_A: p > 0.5$  \* **\*\*Informal Statement\*\*:** The students can correctly identify the pouring order better than by random guessing.  
 isn't the alternative hyp that is it not 0.5, not necessarily greater



You're correct! The alternative hypothesis could indeed be stated as **"not equal to 0.5"** rather than specifically **"greater than 0.5"**, depending on whether we are interested in any deviation from 50% (both higher and lower success rates). This would make it a **two-tailed test** instead of a **one-tailed test**.

Let's clarify the hypotheses accordingly:

## Hypotheses for a Two-Tailed Test:

- **Null Hypothesis**  $H_0$ : The students are guessing randomly, so the probability of correctly identifying the pouring order is exactly 0.5.

$$H_0 : p = 0.5$$

- **Alternative Hypothesis**  $H_A$ : The probability of correctly identifying the pouring order is different from 0.5.

$$H_A : p \neq 0.5$$

- **Informal Statement:** The students have some ability to identify the pouring order that differs from random guessing, either better or worse than 50%.

For a **two-tailed test**, you would modify the binomial test to check for deviations in both directions:

python

```
# Perform the binomial test for a two-tailed hypothesis
result = binomtest(correct_guesses, total_students, probability_under_null,
alternative='two-sided')

# Output the p-value and test statistic
print("P-value:", result.pvalue)
```

This approach tests for any significant difference from 50%, regardless of whether it is higher or lower. It's a more general test if you're only interested in checking for a difference rather than an improvement.

Would you like further clarification or help with running this code?



why cant i do like a bootstrapped sample



You can absolutely use a **bootstrapping** approach to test this hypothesis! Bootstrapping is a powerful method, especially when you want to estimate the sampling distribution of a statistic without making assumptions about its exact distribution. Here's how you can apply bootstrapping to this problem:

## Steps for Bootstrapping Approach

### 1. Define the Hypothesis:

- Similar to before, your null hypothesis is  $H_0 : p = 0.5$ .
- Alternative hypothesis  $H_A : p \neq 0.5$  (for a two-tailed test).

### 2. Generate Bootstrap Samples:

- Under the null hypothesis, each student has a 50% chance of guessing correctly.
- Generate many simulated samples (e.g., 10,000) where each sample is composed of 80 "guesses" with a probability of 0.5 for being correct.

### 3. Calculate the Test Statistic:

- For each bootstrap sample, calculate the proportion of correct guesses.
- Collect these proportions to form the bootstrap distribution of the test statistic.

### 4. Calculate the p-value:

- Compare the observed test statistic (49/80) to the bootstrap distribution.
- For a two-tailed test, calculate the proportion of bootstrap samples that are as extreme or more extreme than the observed proportion in either direction (above or below the observed proportion).

Let's go through the code to perform this bootstrapped hypothesis test.

## Bootstrapping Code Example

python

```
import numpy as np

# Set up parameters
np.random.seed(1) # For reproducibility
total_students = 80
observed_correct_guesses = 49
observed_proportion = observed_correct_guesses / total_students
num_simulations = 10000
null_probability = 0.5

# Generate bootstrap samples under the null hypothesis
bootstrap_samples = np.random.binomial(total_students, null_probability,
size=num_simulations)
bootstrap_proportions = bootstrap_samples / total_students

# Calculate the "as or more extreme" values for a two-tailed test
extreme_values = (abs(bootstrap_proportions - null_probability) >= abs(observed_proportion
- null_probability))
p_value = np.mean(extreme_values)

# Output the p-value
print("Bootstrapped P-value:", p_value)
```

## Explanation of the Code

- We generate `num\_simulations` (10,000) samples where each student has a 50% chance of guessing correctly.
- For each sample, we calculate the proportion of correct guesses.
- Then, we compare how often the bootstrapped proportions are as extreme as or more extreme than our observed proportion (0.6125). This gives us a two-tailed p-value.

## Interpreting the Bootstrapped p-value

- If the p-value is small (typically less than 0.05), it suggests that the observed proportion of correct guesses (49 out of 80) is unlikely to occur by chance, allowing us to reject the null

hypothesis.

- A larger p-value means the observed proportion could plausibly occur due to random guessing.

Using a bootstrapped approach provides a flexible way to estimate the sampling distribution without relying on the exact form of the binomial distribution. This method is useful if you want to visually inspect or further analyze the distribution of sample proportions. Let me know if you'd like any further clarification on this approach!



ok so the p value i got was 0.0445



With a p-value of **0.0445**, this result suggests:

### 1. Interpretation:

- Since the p-value is **less than 0.05**, you have **statistically significant evidence** at the 5% level. This means that under the null hypothesis of random guessing (50% probability), getting a result as extreme as 49 correct guesses out of 80 would be quite unlikely.

### 2. Conclusion:

- You can **reject the null hypothesis** at the 5% significance level. This provides evidence that the students' ability to identify the pouring order differs from random guessing, suggesting that they may indeed have some ability to tell the difference.

### 3. Context:

- While the result is statistically significant, it's important to consider the context of the experiment. A p-value of 0.0445 indicates some evidence against the null hypothesis, but in practical terms, this does not necessarily mean the difference is large or that the students have a high accuracy rate—just that their performance differs from random guessing.

---

This bootstrapping approach provides a robust, flexible method for hypothesis testing, and your p-value interpretation aligns well with standard practice for statistical conclusions. Would you like any further analysis or visualizations to support your findings?



okay and explain to me what we actually did, like simply overall in this experiment, what were the steps we took, what did we conclude and how does this answer the original question



Let's summarize what we did in this experiment, the steps we took, our conclusion, and how it answers the original question:

## What We Did

### 1. Defined the Problem and Hypotheses:

- The question was whether students could identify if milk or tea was poured first better than by random guessing.
- We set up hypotheses:
  - **Null Hypothesis ( $H_0$ ):** The students are just guessing, so the chance of correctly identifying the pouring order is 50% (0.5).
  - **Alternative Hypothesis ( $H_A$ ):** The students' ability to identify the pouring order is different from 50% (could be better or worse).

### 2. Collected Data:

- In this experiment, we had a sample of 80 students, of which 49 correctly identified the pouring order.
- The observed proportion of correct answers was  $\frac{49}{80} = 0.6125$ .

### 3. Used Bootstrapping for Hypothesis Testing:

- We used a bootstrapping approach to simulate what would happen if the students were just guessing.
- We generated 10,000 samples under the assumption that each student had a 50% chance of guessing correctly.
- For each sample, we calculated the proportion of correct answers and formed a distribution to see how extreme our observed proportion (0.6125) was compared to this distribution.

### 4. Calculated the p-value:

- We found how often the simulated proportions were as extreme or more extreme than our observed result in both directions (since this was a two-tailed test).
- The resulting p-value was **0.0445**, which represents the probability of observing a result as extreme as ours if students were truly guessing randomly.

## What We Concluded

- The p-value of 0.0445 is less than the standard significance level of 0.05, so we have **enough evidence to reject the null hypothesis**.
- This means that the observed data (49 out of 80 correct guesses) is unlikely to have occurred by random chance alone. Thus, we conclude that the students' ability to identify the pouring order **differs from random guessing**.

## How This Answers the Original Question

The original question was about determining whether students could differentiate the pouring order by something other than just guessing. By testing the null hypothesis and finding a statistically significant p-value, we showed that the students' performance was likely not due to chance.

- **In simple terms:** The data suggests that the students are not simply guessing; they may have some ability to tell the difference.
- **In statistical terms:** We found evidence that the proportion of correct answers is significantly different from what would be expected under random guessing.

This bootstrapped approach allowed us to assess the likelihood of observing such a result if guessing were the only factor, giving us insight into whether the students' performance could reflect a real skill rather than chance.



#### 4. Quantitative Analysis

Next, what we are going to do is use bootstrapping to find  
i have this explanation that im gonna put in a cell before doing the bootstrapping, how do i word  
what and why we are going to do this>



Here's a concise way to explain what you're doing with bootstrapping and why:

### Quantitative Analysis Explanation

To determine if the students' performance differs significantly from random guessing, we will use **bootstrapping**. Bootstrapping allows us to simulate many samples under the assumption of the null hypothesis (that students are guessing), which provides an estimated distribution of correct guesses if guessing were the only factor involved.

#### 1. What We're Doing:

- We'll simulate a large number of samples (e.g., 10,000) where each student has a 50% chance of correctly guessing whether milk or tea was poured first.
- For each sample, we'll calculate the proportion of correct answers, creating a distribution of these proportions under the null hypothesis.

#### 2. Why We're Doing This:

- This bootstrapped distribution will show us the range of results we might expect if students were purely guessing.
- By comparing our actual observed proportion (49/80) to this distribution, we can calculate a **p-value**—the probability of seeing a result as extreme as ours under the null hypothesis.
- If this p-value is small, it suggests that our observed result is unlikely under random guessing, giving us evidence to reject the null hypothesis.

This explanation helps set up the bootstrapping process by clarifying the purpose (estimating what results look like under random guessing) and how it leads to your conclusion (through the p-value).



thats not as concise as i would like



Here's a more concise version:

## Quantitative Analysis Explanation

To test if students are guessing, we'll use **bootstrapping** to simulate results under the assumption that each student has a 50% chance of guessing correctly. By generating 10,000 simulated samples, we'll create a distribution of correct answers expected by chance alone. This lets us calculate a **p-value**, showing how likely our observed result (49 out of 80) is if students were only guessing. A low p-value indicates our result is unlikely under the null hypothesis, suggesting that students may be doing better than random guessing.

This should set up the analysis effectively while remaining brief!



Quantitative analysis addressing the validity of  
Explanation of the method clearly articulating the purpose of the usage of statistic(s) to address  
the population parameter of interest  
i feel as this is lacking in my assignment



To strengthen this section of your assignment, you'll want to clearly outline the **purpose of your analysis** and **why you chose specific statistical methods** in the context of the population parameter of interest. Here's a more detailed structure for this part:

## Quantitative Analysis Addressing the Validity of the Null Hypothesis

In this experiment, our population parameter of interest is the **proportion of students who can correctly identify whether milk or tea was poured first**. The null hypothesis assumes that students are guessing, meaning this proportion is 0.5 (50%).

To assess this, we use **bootstrapping** because it allows us to simulate many possible outcomes under the null hypothesis without assuming a specific distribution. By repeatedly sampling with the null hypothesis of 50% correct guesses, we create a distribution of sample proportions that represents what we would expect if students were purely guessing.

## Explanation of the Method

### 1. Purpose of Bootstrapping:

- Bootstrapping is a resampling technique that approximates the sampling distribution of a statistic—in this case, the proportion of correct guesses. By simulating thousands of trials, bootstrapping helps us understand the range of results we might expect due to chance.

### 2. Calculating the p-value:

- We compare the observed proportion of correct answers ( $49/80 = 0.6125$ ) to our bootstrapped distribution. The **p-value** we calculate reflects the probability of obtaining a



result as extreme as or more extreme than our observed proportion, assuming the null hypothesis is true.

### 3. Addressing the Population Parameter:

- The bootstrapping approach lets us empirically estimate the distribution of correct responses under the null hypothesis for our sample. By examining where our observed statistic falls within this distribution, we can determine if there's evidence to suggest the true population proportion is different from 0.5.

## Why This Approach is Effective

This method is particularly useful when working with proportions and allows us to draw conclusions about the **population parameter** of interest without heavily relying on theoretical assumptions. A small p-value from this process would imply that the observed performance of students (49 out of 80) is unlikely to be due to random guessing, indicating potential evidence against the null hypothesis.

This format addresses why bootstrapping is appropriate for your experiment and how it directly relates to understanding the population parameter, enhancing the rigor of your quantitative analysis section.



what is the population parameter, and what kind of conclusions does bootstrapping and the sample mean and all that give or not give about it



In this context, the **population parameter** of interest is the **true proportion** of students who can correctly identify whether milk or tea was poured first. In other words, it represents the probability that any given STA130 student (from the larger population) can correctly distinguish the pouring order.

## What Bootstrapping Tells Us About the Population Parameter

Bootstrapping helps us understand the behavior of the **sample statistic** (here, the proportion of correct answers) under the assumption of a particular population parameter (in this case, 0.5 under the null hypothesis). By simulating many samples, we can observe how much variation we'd expect in the proportion of correct answers purely by chance. Here's what bootstrapping can and cannot tell us about the population parameter:

### What Bootstrapping Can Tell Us:

#### 1. Sampling Distribution of the Sample Statistic:

- Bootstrapping approximates the sampling distribution of the sample mean or proportion under the null hypothesis. This gives us a sense of the range of sample statistics we might observe if the population parameter were truly 0.5.

#### 2. p-value:

- By comparing our observed sample proportion ( $49/80 = 0.6125$ ) to the bootstrapped distribution, we can calculate a p-value. This tells us the probability of seeing a result as extreme as or more extreme than our observed value, assuming the population parameter is indeed 0.5.
- A low p-value suggests that the observed sample proportion is unlikely under the null hypothesis, which provides evidence against it.

### 3. Confidence Intervals (if used):

- Bootstrapping can also create confidence intervals for the sample proportion, which estimate the range in which we believe the true population parameter likely falls. For example, a 95% confidence interval provides a range where we would expect the true proportion to fall 95% of the time based on our sample data.

## What Bootstrapping Cannot Tell Us:

### 1. Exact Value of the Population Parameter:

- Bootstrapping does not give the exact value of the population parameter. Instead, it gives an estimate based on the sample data. We can't say with certainty what the true proportion is; we can only describe the likely range based on our sample.

### 2. Causation or Mechanisms:

- Bootstrapping can tell us whether our sample proportion is unusual under the null hypothesis, but it doesn't explain *why* students are able to distinguish the pouring order. It only provides evidence about the likelihood of random guessing as an explanation.

### 3. Deterministic Proof of the Population Parameter:

- Bootstrapping provides statistical evidence, not absolute proof. Even with a small p-value, we cannot say with certainty that students truly have the ability to distinguish the pouring order—it's always possible (though unlikely) that the sample proportion was due to chance.

## Conclusion

Bootstrapping, along with the observed sample proportion, gives us a way to infer whether the true population parameter is likely different from the hypothesized value (0.5 under the null). It provides evidence **against** the null hypothesis if the sample proportion is significantly different, but it doesn't provide the exact population parameter or conclusive proof. Instead, it allows us to make probabilistic statements about the likelihood of our sample results occurring by chance, helping us to understand and draw conclusions about the population parameter indirectly.



if i got heads 8 out of ten times i flipped a coin, what would the p value likely be, like large or small



If you got **heads 8 out of 10 times** when flipping a fair coin, the p-value would likely be **small**. Here's why:

## Explanation

Under the null hypothesis, a fair coin has a 50% chance of landing heads on any given flip. If we assume the coin is fair, the expected number of heads in 10 flips would be 5 (or close to 5 with some variation).

Getting **8 heads out of 10 flips** is an extreme result when compared to this expectation. You can use a **binomial test** to determine the exact p-value, but intuitively, here's what you're looking at:

### 1. Likelihood of 8 Heads:

- Flipping a coin 10 times and getting 8 heads is relatively rare if the true probability of heads is 0.5.

### 2. Two-Tailed Test:

- Since you're testing if the number of heads is different from 5 in either direction, you'd consider extreme results on both sides. However, with 8 heads, you're particularly interested in values much higher than 5.

### 3. Resulting p-value:

- Because 8 heads deviates quite a bit from the expected 5 heads, the p-value for this outcome would be **small**. A small p-value indicates that such an extreme result is unlikely to happen by random chance alone, suggesting that the observed result is not typical of a fair coin.

## Conclusion

In summary, getting 8 heads out of 10 flips would likely result in a **small p-value**, indicating that the outcome is unusual for a fair coin. This might lead you to question the fairness of the coin, especially if you repeatedly observe such an extreme result.



"It is important to note that outcomes of tests refer to the population parameter, rather than the sample statistic! As such, the result that we get is for the population." In terms of the distinctions between the concepts of 's, , and , how would you describe what the sentence above means? Explain this concisely in your own words for a "non-statistical" audience,



For a non-statistical audience, here's a simplified explanation of that sentence:

When we run a test, we're not just interested in what happened in our **sample** (the group we actually observed or tested). Instead, we're trying to learn something about the **larger population** from which that sample was taken.

- The **sample statistic** is just a number that describes what we saw in our specific group, like the percentage of people who guessed correctly in our sample.
- But the **population parameter** is what we're really after. This is an estimate of what's true for the entire population, not just our sample.

So, the result of our test helps us make conclusions about the entire population, even though we only collected data from a smaller group. In other words, we're using our sample to draw a broader conclusion.



the test statistic is what you observed from the sample, and the population mean is what the null hypothesis suggests a value for. the p-value is the probability that ( assuming the null hypothesis about the population mean was true) you would observe an as or more extreme test statistic then the one you already have, if you sampled again independently using the same sample szie.

0

can u explain this simply



Absolutely, here's a simpler version:

- The **test statistic** is just the result you got from your sample (like the number or proportion of correct guesses in your group).
- The **population mean** is what we assume for the whole group if the null hypothesis is true (for example, a 50% chance of guessing correctly for everyone).
- The **p-value** is the chance of seeing a result as extreme as yours (or even more extreme) if you repeated the test with the same sample size, assuming the null hypothesis is correct.

So, a small p-value means that your result is unusual under the null hypothesis, suggesting it might not just be due to random chance.



 Uploaded image

Essentially what this statement is saying is that the tests we perform, and the conclusions we make are geared toward the larger idea of the entire population. The sample statistic that we get is basically just telling us what we specifically observed in our sample. Like the number of times a coin flipped to tails. On the other hand, the population parameter is basically a conclusion that is true for the entire population, allowing us to make more nuanced and accurate conclusions.

does this answer make sense for 2 or should i change stuff



Your answer does make sense, but it could be refined to more clearly distinguish the roles of the sample statistic and the population parameter, and how they relate to hypothesis testing. Here's a revised version that builds on your explanation:

The statement emphasizes that the tests we perform aim to make conclusions about the **population as a whole**, not just the specific group (or sample) we observed. The **sample statistic** is what we directly measure from our sample, like the number of times a coin landed on tails. This value describes only what happened in that particular sample.

In contrast, the **population parameter** represents a broader truth we're trying to infer for everyone or everything that could be sampled (like the true probability of getting tails in all coin flips). Hypothesis tests use the sample statistic to estimate the population parameter, allowing us to make more general conclusions about the entire population, rather than just the specific sample we observed.

---

This revised explanation should clarify the distinction and why our tests focus on population parameters, making it clear for a non-statistical audience.



Hypothesis tests use the sample statistic to estimate the population parameter, allowing us to make more general conclusions about the entire population, rather than just the specific sample we observed.

gotta dumb this down i think it needs to be easily understandable



Let's simplify this explanation even more:

---

The tests we do are not just about the specific group we looked at—they help us understand the **whole population**. The **sample statistic** is just what we saw in our sample, like how many times a coin landed on tails. But the **population parameter** is what we believe is true for everyone, not just our sample.

By using the sample statistic, we can make an educated guess about what's likely true for the whole population, so we're not just talking about what happened with our sample but what's probably true overall.

---

This version keeps it very straightforward, focusing on the main idea without technical terms.



give me a summary of our conversation



Here's a summary of our conversation:

### 1. Assignment Assistance:

- You asked for help with various statistical concepts related to your assignment, including hypothesis testing, p-values, and bootstrapping.

### 2. Explaining Concepts Simply:

- We discussed how to explain statistical ideas in a way that's understandable for a non-statistical audience, like distinguishing between sample statistics and population parameters, and explaining the meaning of a p-value in simple terms.

### 3. Quantitative Analysis Section:

- You were working on a hypothesis test with bootstrapping and wanted to clarify how to describe the purpose and steps. I suggested ways to explain what you're doing with bootstrapping, why it's effective, and how it relates to testing the null hypothesis about the population parameter.

### 4. Bootstrapping Example:

- We went through a bootstrapping approach to test whether a sample proportion (e.g., students guessing correctly) is significantly different from random guessing. You calculated a p-value of 0.0445, indicating statistically significant results against the null hypothesis.

### 5. Simplifying Explanations:

- We worked on simplifying your explanations for parts of the assignment, like defining sample statistics vs. population parameters, so they're more accessible for a general audience.

### 6. Further Clarifications:

- You shared some draft answers, and I provided suggestions to make them clearer and more concise, especially for non-technical readers.

If there's anything more specific or another part of the conversation you'd like detailed, just let me know!

ChatGPT can make mistakes. Check important info.